Genetic Programming for Evolutionary Feature Construction in Regression

Hengzhe Zhang

Supervisor: Mengjie Zhang, Bing Xue, Qi Chen

Victoria University of Wellington

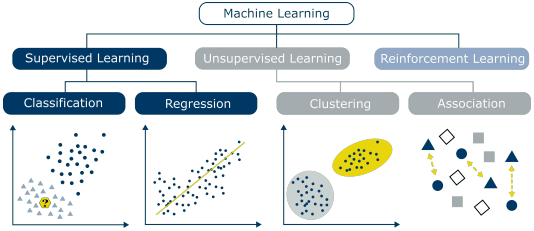
30/08/2023

Table of Contents

Introduction

Machine Learning

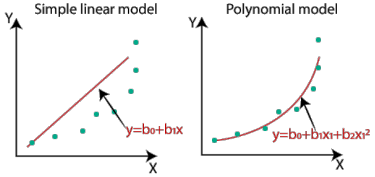
- A computational method that uses past experiences to generate accurate predictions for future data.
- Applications: disease diagnosis, fraud detection, and natural disaster prediction, etc.



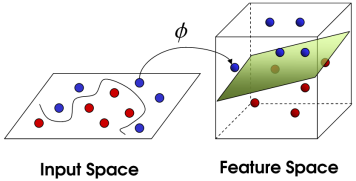
Machine Learning

Feature Construction

- The general idea of feature construction is to construct a set of new features $\{\phi_1, \ldots, \phi_m\}$ to enhance the learning performance of machine learning algorithms on a given dataset $\{\{x_1, y_1\}, \ldots, \{x_n, y_n\}\}$ compared to learning on the original features $\{x^1, \ldots, x^p\}$.



(a) Feature Construction on Regression



(b) Feature Construction on Classification

An Example

- Body Mass Index (BMI) is an example of a feature constructed by combining weight and height measurements.
- BMI is associated with obesity levels in patients from New Zealand.

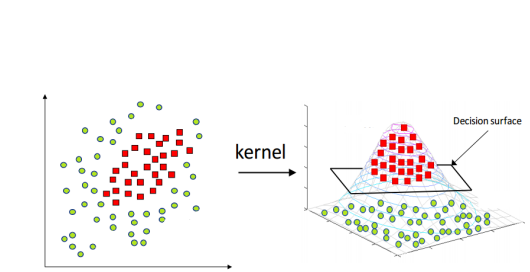
$$\text{BMI} = \frac{\text{Weight (in kilograms)}}{\text{Height}^2 \text{ (in meters)}}$$

Body Mass Index

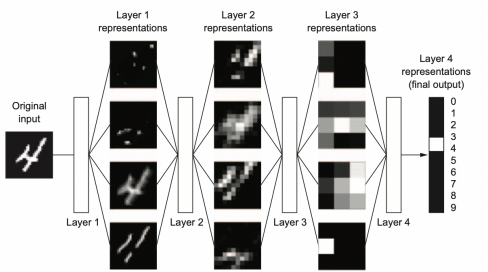
[1]Mohammed A Moharram et al. (2020). "Correlation between epicardial adipose tissue and body mass index in New Zealand ethnic populations". In: *The New Zealand Medical Journal (Online)* 133.1516, pp. 22–5.

Feature Construction Methods

- Kernel Methods
- Deep Learning



(a) Kernel Methods



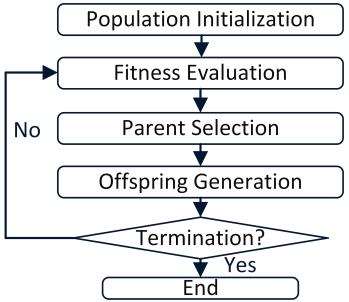
(b) Deep Learning

Genetic Programming

- Genetic programming is a population-based search method with variable-length representation that aims to search for a computer program capable of solving a given task.

Advantages:

- Flexible Length of Representation
- Population-Based Search Algorithm
- Gradient-Free Search Mechanism
- Good Interpretability

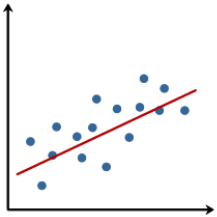


The evolution process of GP.

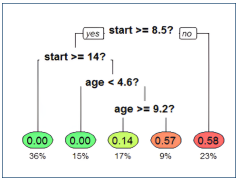
Fundamental Research

Feature Construction:

- For an ensemble of linear models (CIM 2023)[1]
- For an ensemble of decision trees (TEVC 2021)[2]



(a) Linear Model



(b) Decision Tree

[1] Hengzhe Zhang, Qi Chen, et al. (2023). "MAP-Elites for Genetic Programming-based Ensemble Learning: An Interactive Approach". In: *IEEE Comput. Intell. Mag.*

[2] Hengzhe Zhang, Aimin Zhou, and Hu Zhang (2022). "An Evolutionary Forest for Regression". In: *IEEE Trans. Evol. Comput.* 26.4, pp. 735–749.

Feature Construction:

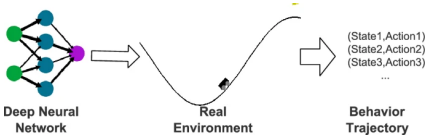
- For an ensemble of linear models and decision trees (TEVC 2023)
- Significantly better than state-of-the-art ensemble learning methods such as Random Forest, XGBoost, and LightGBM.

Linear Models and Decision Trees

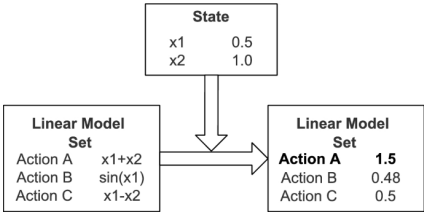
[1] Hengzhe Zhang, Aimin Zhou, Qi Chen, et al. (2023). "SR-Forest: A Genetic Programming based Heterogeneous Ensemble Learning Method". In: *IEEE Trans. Evol. Comput.*

Feature Construction:

- For reinforcement learning/imitation learning (Comp. Intell. Syst. 2020)[1]
- Through feature construction, a linear model can achieve performance comparable to that of deep neural networks.



(a) Data Collection

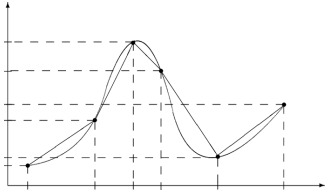


(b) Imitation Learning

[1] Hengzhe Zhang, Aimin Zhou, and Xin Lin (2020). "Interpretable policy derivation for reinforcement learning based on evolutionary feature synthesis". In: *Comp. Intell. Syst.* 6, pp. 741–753.

Feature Construction:

- For piecewise regression (SWEVO 2022)
- Significantly better than state-of-the-art symbolic regression methods such as FEAT and Operon.



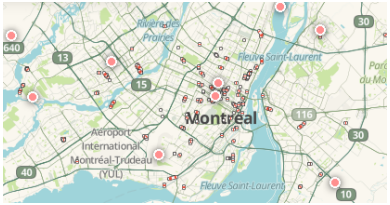
Piecewise Regression

[1] Hengzhe Zhang, Aimin Zhou, Hong Qian, et al. (2022). "PS-Tree: A piecewise symbolic regression tree". In: *Swarm Evol. Comput.* 71, p. 101061.

Applied Research

Bike lane usage forecasting

- Bike lane usage forecasting for Montreal, Canada, which provides valuable insights for urban planning and the optimization of transportation systems.
- Given a specific lane and associated weather conditions, our feature construction technique can be employed to develop a learning model f that is capable of predicting the usage of that lane on a given date.
- Winner of GECCO 2023 "Interpretable Symbolic Regression for Data Science" competition.



Map of Montreal

Conclusion

Conclusion

- Evolutionary feature construction is an effective technique for improving learning performance.

Future Directions

- Applying feature construction to real-world applications, including but not limited to cyber-marine seafood, climate change, and marine genomics.

Thanks for listening!

Email: Hengzhe.zhang@ecs.vuw.ac.nz